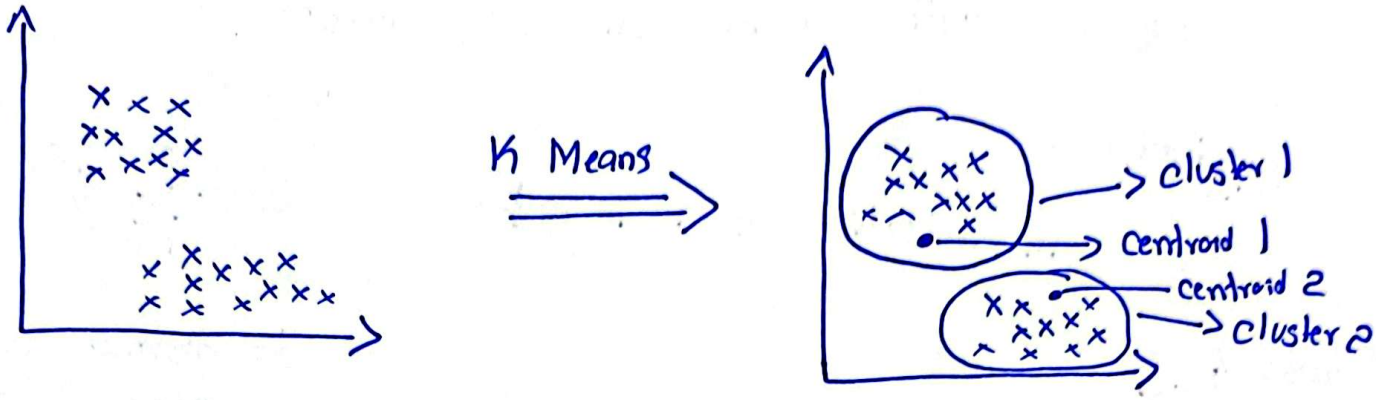
K Means Clustering



Steps:

1. Decide how many clusters you want i.e., choose K
2. Randomly assigned a centroid to each of the K cluster
3. Calculate the distance of all observation to each of the K centroids.
4. Assign observations to the closet centroid
5. Find the new location of the centroid by taking mean of all observations in each cluster.
6. Repeat steps 3-5 until the centroids do not change position

⟶ How to find K-values

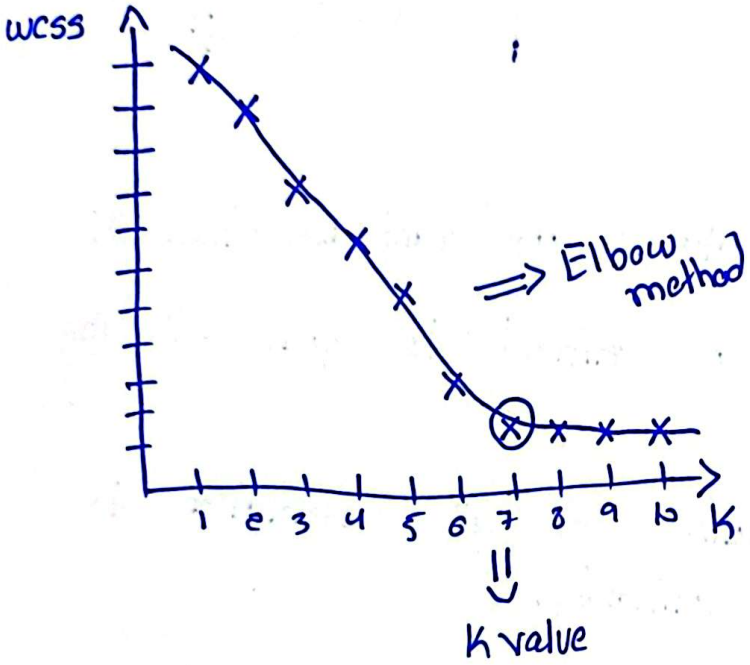
WCSS = with cluster sum of squares

Suppose:

Initialize K = 1 to 10

$$WCSS = \sum_{i=1}^{n} d^2$$

d = distance between points to nearest centroid



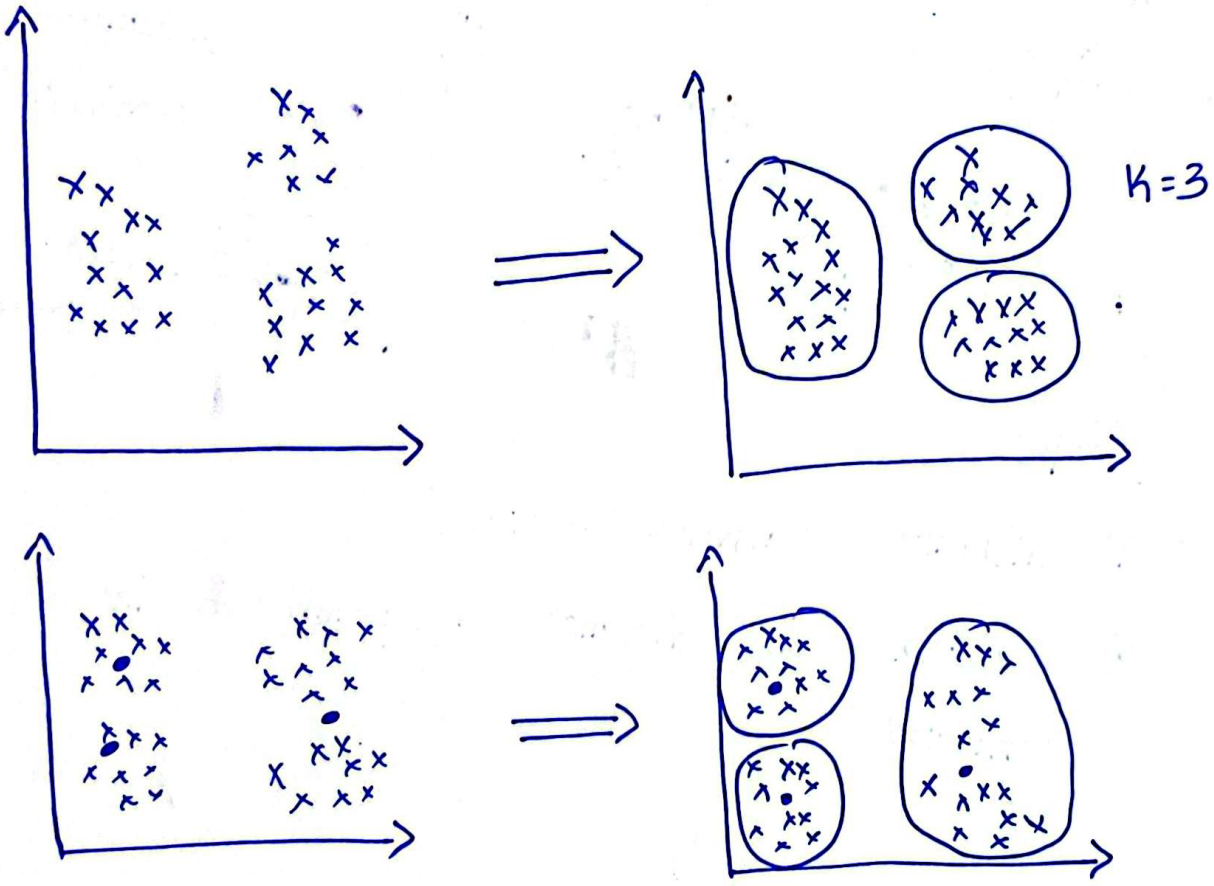
① Euclidean distance

$$\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$$

② Manhattan distance

$$|x_1-x_2|+|y_1-y_2|$$

⟶ Random Initialization Trap (K Means ++)



K Means ++ Initialization

It carefully choose centroids to be as spread out as possible.